



PATENT ABSTRACTS OF JAPAN

(11) Publication number: 09171479 A

(43) Date of publication of application: 30.06.97

(51) Int. Cl

G06F 12/00**G06F 12/16**

(21) Application number: 07330765

(71) Applicant: NEC CORP

(22) Date of filing: 19.12.95

(72) Inventor: KIKUCHI YOSHIHIDE

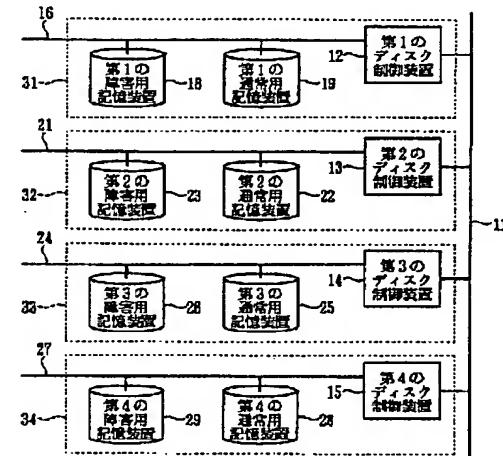
(54) DATA STORAGE SYSTEM

COPYRIGHT: (C)1997,JPO

(57) Abstract:

PROBLEM TO BE SOLVED: To provide a data storage system with which the increase of burden on other disk control device for substituting reading and writing can be reduced when any fault is generated at a disk control device for controlling the read/ write of disk device.

SOLUTION: The same data as data stored in an ordinary storage device 19 of 1st unit server 31 are distributedly stored in storage devices 23, 26 and 29 for fault of other unit servers 22, 25 and 28. Similarly, the copies of data stored in the ordinary storage devices of respective servers 32-34 are distributedly prepared in the storage devices for fault of other unit servers. When any fault occurs at the ordinary storage device of one unit server or disk control devices 12-15, the data distributed to the storage devices for fault of the other plural unit servers are read out. Since the data in the storage device, where the fault occurs, are distributed to the other plural servers, the burden on disk control devices of the other unit servers is not remarkably increased.



(51) Int.Cl. ⁵ G 0 6 F 12/00 12/16	識別記号 5 3 1 3 1 0	序内整理番号 7623-5B	F I G 0 6 F 12/00 12/16	技術表示箇所 5 3 1 D 3 1 0 M
---	------------------------	-------------------	-------------------------------	------------------------------

審査請求 有 請求項の数 5 O.L (全 15 頁)

(21)出願番号 特願平7-330765	(71)出願人 000004237 日本電気株式会社 東京都港区芝五丁目7番1号
(22)出願日 平成7年(1995)12月19日	(72)発明者 菊地 芳秀 東京都港区芝五丁目7番1号 日本電気株式会社内

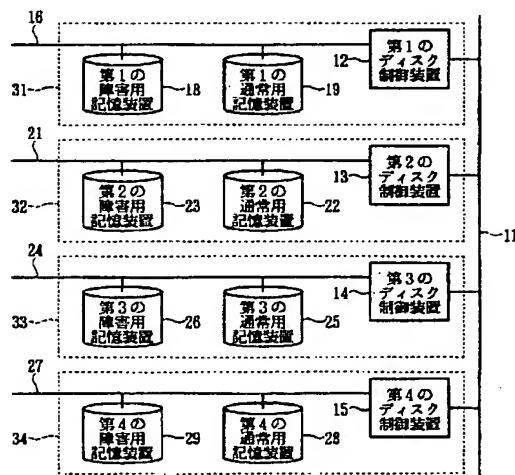
(74)代理人 弁理士 山内 梅雄

(54)【発明の名称】 データ格納システム

(57)【要約】

【課題】ディスク装置の読み書きを制御するディスク制御装置に障害が生じとき、これに代わって読み書きを行う他のディスク制御装置の負担の増加を軽減することのできるデータ格納システムを提供する。

【解決手段】第1単位サーバ31の通常用記憶装置17に格納されているデータと同一のデータを、他の単位サーバ22、25、28の障害用記憶装置23、26、29に分散して格納する。同様した各単位サーバの通常用記憶装置に格納されているデータの複製を他の単位サーバの障害用記憶装置に分散して作成する。1つの単位サーバの通常用記憶装置またはディスク制御装置12～15に障害が起きたときは、他の複数の単位サーバの障害用記憶装置に分散されたデータを読み出す。障害の起きた記憶装置のデータは他の複数のサーバに分散されているので、他の単位サーバのディスク制御装置の負担が大幅に増加しない。



【特許請求の範囲】

【請求項1】 データを蓄積するための第1および第2のデータ記憶手段と、これらデータ記憶手段に入出力すべきデータを所定のネットワークとの間で送受信する通信手段と、この通信手段により前記ネットワークを通じて前記第1のデータ記憶手段に記憶すべきデータを受信したときこれを第1のデータ記憶手段に書き込む第1の書き込み手段と、この第1の書き込み手段によって前記第1のデータ記憶手段にデータが書き込まれたときこれと同一のデータを前記ネットワークを通じて予め定められた複数の転送先に分散して送出する分散送出手段と、前記ネットワークを通じて他の装置の前記分散送出手段から送出されたデータを受信したときこれを第2のデータ記憶手段に書き込む第2の書き込み手段とをそれぞれ有する複数の単位サーバと、
これら単位サーバの障害の有無を検出する障害検出手段と、
この障害検出手段によって障害の有ることが検出された単位サーバの有する前記第1のデータ記憶手段に記憶されているデータを読み出すときこれと同一のデータをその第2のデータ記憶手段に格納している単位サーバから対応するデータを読み出す障害用読み出手段とを具備することを特徴とするデータ格納システム。

【請求項2】 前記第1のデータ記憶手段にはファイルを単位としてデータが書き込まれ、前記分散送出手段は、第1のデータ記憶手段に書き込まれたファイルを単位として前記予め定められた複数の転送先に第1のデータ記憶手段の記憶内容を分散させることを特徴とする請求項1記載のデータ格納システム。

【請求項3】 前記第1のデータ記憶手段の記憶領域は複数のブロックに分割されており、前記分散送出手段はこのブロックを単位として前記予め定められた複数の転送先に第1のデータ記憶手段の記憶内容を分散させることを特徴とする請求項1記載のデータ格納システム。

【請求項4】 前記複数の単位サーバは、それぞれ複数の単位サーバから構成された2以上のグループに分割されており、前記予め定められた複数の転送先は各グループ内における他の単位サーバであることを特徴とする請求項1記載のデータ格納システム。

【請求項5】 前記複数の単位サーバは、それぞれ複数の単位サーバから構成された2以上のグループに分割されており、前記予め定められた複数の転送先はそれぞれ他のグループの単位サーバであることを特徴とする請求項1記載のデータ格納システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、ディスク装置などの記憶装置を用いて各種データを格納するデータ格納システムに係わり、特に1つの記憶装置に記憶されているデータの複製を他の記憶装置に記憶することによりシス

テムの障害に備えるデータ格納システムに関する。

【0002】

【従来の技術】 ディスク装置などの記憶装置にデータを格納するシステムでは、記憶装置に障害が発生したときでも格納されているデータが失われないようにデータを冗長構成にして記憶することが行われる。従来、ディスク装置の障害対策として、データを冗長構成にして複数のディスク装置に格納するものには、R A I D (Redundant and Array of Inexpensive Disks) と呼ばれるものがある。R A I Dはレベル1からレベル5までが一般的に知られている。

【0003】 レベル1は、通常運用されるディスク装置に格納する内容と同一の内容を障害用のディスク装置に格納するものであり、ミラー方式と呼ばれている。ミラー方式のデータ格納システムでは、通常運用するための通常用ディスク装置と、障害対策用のミラーディスク装置を備えている。ホストコンピュータからデータを書き込むときは、通常用ディスク装置と同一のデータをミラーディスク装置にも格納するようになっている。通常用ディスク装置に障害が生じたときは、ミラーディスク装置から読み出すことにより、データが読み出せなくなるという事態を回避している。

【0004】 レベル2～レベル5では、パリティ情報を生成し、1つのディスク装置に障害が生じたとき他のディスク装置とパリティ情報を用いて障害の生じたディスク装置に格納されている内容を復元するものである。

【0005】 特開平-318640号公報には、パリティ情報を用いたR A I D方式によって障害の生じたディスク装置の記憶内容を復元するデータ格納システムが開示されている。パリティ情報を用いる場合、通常は、たとえば4台のディスク装置のうちの3台のディスク装置にデータを格納し、残りのディスク装置に他の3台のディスク装置に格納したデータに対応するパリティ情報を格納する。このようにデータとそのパリティ情報を格納しているディスク装置をE C Cグループと呼ぶことにする。データを格納している3台のうちの1台が故障したときは、残り2台の正常なディスク装置のデータとパリティ情報を基にして障害の生じたディスク装置のデータが復元される。

【0006】 ディスク装置の台数が8台の場合には、これを4台ずつのE C Cグループに分けると、障害の生じたディスク装置のデータを復元するための負荷がそのE C Cグループ内に集中し、効率良く復元作業を行うことができない。そこで特開平-318640号に開示されている先行技術では、各ディスク装置の記憶領域を複数のブロックに分割し、ブロックごとに互いに異なる4台のディスク装置によってE C Cグループを形成するようにしている。これにより1台分のディスク装置のデータを復元する際の負荷を残りの7台のディスク装置に分散することができる。

【0007】バリティ情報を用いたデータ格納システムでは、ディスク装置の記憶内容を更新するごとにバリティ情報を計算して求めなければならず、高速な書き込み処理を行うことが難しい。またディスク装置に障害が生じたとき、そのデータの復元作業を行う必要があり、データが読み出されるまでの時間が長くかかり、高速な読み出しに対応できない場合がある。これに対してミラー方式では、冗長度が大きいので障害対策用に用意すべきディスク装置の容量が大きくなるが、高速なアクセスが可能になるという利点がある。

【0008】図8は、従来から使用されているミラー方式を用いたデータ格納システムの構成の概要を表したものである。通常時に運用されるディスク装置301と障害対策用のディスク装置302はバス303を通じてディスク制御装置304にそれぞれ接続されている。ディスク制御装置304は、バスを通じて各ディスク装置とデータを入出力するための図示しないインターフェイス回路と、データの読み書きを制御するためのこれまた図示しないCPU(中央処理装置)を備えている。ディスク制御装置304から通常用ディスク装置301へ書き込み命令が出されると、書き込まれるデータは、通常用ディスク装置301に格納されると同時に障害用ディスク装置302にも書き込まれる。

【0009】このように同一のデータを2つのディスク装置に書き込むことにより、通常用ディスク装置301の内容と障害用ディスク装置302の記憶内容は同一になる。通常用ディスク装置301と同一内容のデータを障害用ディスク装置302が保持していることから、ディスク装置302はミラーディスクと呼ばれている。通常用ディスク装置301に障害が生じると、ディスク制御装置304は障害用ディスク装置302からデータを読み出すことで障害に対応するようになっている。

【0010】図8に示したデータ格納システムは、ディスク装置で生じる障害を対象としたものであるが、障害がディスク装置へのデータの読み書きを制御するディスク制御装置に発生することにより、通常用および障害用の双方のディスク装置の内容が読み出せなくなる場合もある。

【0011】図9は、ディスク制御装置の障害に対応することのできるデータ格納システムの構成の概要を表したものである。このシステムでは、第1の通常用ディスク装置311および第1の障害用ディスク装置312はバス313によって第1のディスク制御装置314と接続されている。また、ディスク装置311、312はバス315を通じて第2のディスク制御装置316にも接続されている。同様に、第2の通常用ディスク装置321および第2の障害用ディスク装置322はバス323によって第2のディスク制御装置316に接続されるとともに、バス324により第1のディスク制御装置314にも接続されている。また、第1および第2のディ

スク制御装置314、316はネットワーク325を介して互いに接続されている。さらにネットワーク325には、データの読み出し要求や書き込み要求を行うホストコンピュータ326が接続されている。ディスク制御装置はサーバとして、またホストコンピュータはクライアントとして機能する。

【0012】第1のディスク制御装置314によってデータを第1の通常用ディスク装置311に書き込む際、同じ内容のデータが第1のディスク制御装置314により第1の障害用ディスク装置312にも書き込まれる。

同様に第2のディスク制御装置316によってデータを第2の通常用ディスク装置321に書き込む際、同じ内容のデータが第2のディスク制御装置316により第2の障害用ディスク装置323にも書き込まれる。このようにして、第1の通常用ディスク装置311と第1の障害用ディスク装置312は互いに同じ内容を保持する。また第2の通常用ディスク装置321と第2の障害用ディスク装置322も互いに同じ内容を保持するようになっている。

【0013】今、ネットワーク325に接続されている統括管理用のホストコンピュータ326から第1の通常用ディスク装置311に格納されているデータの読み出し要求があったものとする。システムが正常なときは第1の通常用ディスク装置311の内容はバス313を通じて第1のディスク制御装置314によって読み出され、ネットワーク325を通じてホストコンピュータ326に転送される。第1の通常用ディスク装置311に障害が発生しているときは、第1のディスク制御装置314は第1の通常用ディスク装置311と同一の内容を格納している第1の障害用ディスク装置312から読み出し要求のあったデータをバス313を通じて読み出す。そしてこれをネットワーク325を介してホストコンピュータ326に転送する。

【0014】一方、第1のディスク制御装置314に障害が起きた場合は、第1の通常用ディスク装置311の内容はバス315を通じて第2のディスク制御装置316から読み出され、ネットワーク325を介してホストコンピュータ326に転送される。このようにしてディスク制御装置に障害が発生した場合でも、格納されているデータを読み出すことができる。このシステムと同様にディスク装置の障害とディスク制御装置の障害の双方に対応することができ、かつバスの構成をより簡易にしたデータ格納システムもある。

【0015】図10は、ディスク装置の障害およびディスク制御装置の障害の双方に対応することのできるバス構成の簡易なデータ格納システムの概要を表したものである。第1のディスク制御装置331には、第1の通常用ディスク装置332と第1の障害用ディスク装置333がバス334を通じて接続されている。また、第2のディスク制御装置341には、第2の通常用ディスク

装置342と第2の障害用ディスク装置343がバス344を通じて接続されている。第1のディスク制御装置331と第2のディスク制御装置341は互いにネットワーク351を通じて接続されている。さらにネットワーク351には、データの読み出しおよび書き込み要求を行うホストコンピュータ352が接続されている。

【0016】このシステムでは、第1のディスク制御装置331に接続されている第1の通常用ディスク装置332に対応するミラーディスク装置は、第2のディスク制御装置341のバス344に接続されている第2の障害用ディスク装置343を用いる。また第2のディスク制御装置341に接続されている第2の通常用ディスク装置342に対応するミラーディスク装置は、第1のディスク制御装置331に接続されている第1の障害用ディスク装置333を用いる。

【0017】第1の通常用ディスク装置332に書き込むデータと同一のデータは、ネットワーク351および第2のディスク制御装置341を通じて第2の障害用ディスク装置343にも書き込まれる。また、第2の通常用ディスク装置342に書き込むデータと同一のデータは、第1のディスク制御装置331を通じて第1の障害用ディスク装置333にも書き込まれる。第1のディスク制御装置331に障害が発生したときは、第1の通常用ディスク332の内容は第2の障害用ディスク装置343にも格納されているので、第2のディスク制御装置341を通じて第2の障害用ディスク装置から読み出すようになっている。

【0018】

【発明が解決しようとする課題】このように1つのディスク制御装置に接続されている通常運用されるディスク装置の内容を他の1つのディスク制御装置に接続されている障害用のディスク装置に複製しておけば、ディスク装置とディスク制御装置のいずれに障害が発生してもデータの読み出しが行うことができる。しかしながら、ディスク制御装置に障害が発生した場合は、障害の起きたディスク制御装置に接続されているディスク装置に対応するミラーディスク装置を有する他のディスク制御装置の負荷が倍になってしまうという問題がある。図10の例では、第1のディスク制御装置に障害が生じると、第1の通常用ディスク装置の内容はこれと同一内容を保持している第2の障害用ディスク装置から第2のディスク制御装置を通じて読み出すことになる。第2のディスク制御装置は第2の通常用ディスク装置の内容を読み出す役割も負っているため、これら2台分の読み出しを行なわなければならずその負荷が倍になる。その結果、データの読み出し処理に時間がかかるという問題がある。

【0019】そこで本発明の目的は、ディスク装置の読み書きを制御するディスク制御装置に障害が生じたとき、これに代わって読み書きを行う他のディスク制御装置の負担の増加を軽減することのできるデータ格納システム

を提供することにある。

【0020】

【課題を解決するための手段】請求項1記載の発明では、データを蓄積するための第1および第2のデータ記憶手段と、これらデータ記憶手段に入出力すべきデータを所定のネットワークとの間で送受信する通信手段と、この通信手段によりネットワークを通じて第1のデータ記憶手段に記憶すべきデータを受信したときこれを第1のデータ記憶手段に書き込む第1の書き込み手段と、この第1の書き込み手段によって第1のデータ記憶手段にデータが書き込まれたときこれと同一のデータをネットワークを通じて予め定められた複数の転送先に分散して送出する分散送出手段と、ネットワークを通じて他の装置の分散送出手段から送出されたデータを受信したときこれを第2のデータ記憶手段に書き込む第2の書き込み手段とをそれぞれ有する複数の単位サーバと、これら単位サーバの障害の有無を検出する障害検出手段と、この障害検出手段によって障害の有ることが検出された単位サーバの有する第1のデータ記憶手段に記憶されているデータを読み出すときこれと同一のデータをその第2のデータ記憶手段に格納している単位サーバから対応するデータを読み出す障害用読出手段とをデータ格納システムに具備させている。

【0021】すなわち請求項1記載の発明では、通常の運用で用いられる第1のデータ記憶装置に記憶された内容を他の複数の単位サーバの第2のデータ記憶装置に分散して記憶している。そして、障害の有る単位サーバの第1のデータ記憶手段に格納されているデータを読み出すとき、これと同一のデータをその第2のデータ記憶手段に格納している単位サーバから、必要なデータを読み出している。第1のデータ記憶手段の記憶内容が他の複数の単位サーバの第2のデータ記憶手段に分散して格納されているので、1つの単位サーバに障害が起きたときでも、他の単位サーバの負担が大幅に増大するがない。

【0022】請求項2記載の発明では、第1のデータ記憶手段にはファイルを単位としてデータが書き込まれ、分散送出手段は、第1のデータ記憶手段に書き込まれたファイルを単位として予め定められた複数の転送先に第1のデータ記憶手段の記憶内容を分散させている。

【0023】すなわち請求項2記載の発明では、第1のデータ記憶手段の記憶内容をファイル単位に他の複数の単位サーバに分散している。

【0024】請求項3記載の発明では、第1のデータ記憶手段の記憶領域は複数のブロックに分割されており、分散送出手段はこのブロックを単位として予め定められた複数の転送先に第1のデータ記憶手段の記憶内容を分散させている。

【0025】すなわち請求項3記載の発明では、第1のデータ記憶手段の記憶領域を複数のブロックに分割し、

このブロックを単位の第1のデータ記憶手段の記憶内容を複数の単位サーバに分散して格納している。

【0026】請求項4記載の発明では、複数の単位サーバは、それぞれ複数の単位サーバから構成された2以上のグループに分割されており、予め定められた複数の転送先は各グループ内における他の単位サーバに設定されている。

【0027】すなわち請求項4記載の発明では、各単位サーバの有する第1のデータ記憶手段に記憶されたものと同一のデータは、その単位サーバの属するグループ内における他の複数の単位サーバの第2のデータ記憶手段に分散されて格納される。これにより、各グループ内で1つの単位サーバの障害をリカバリーすることができる。データ格納システム全体として2以上の単位サーバの障害に対応することができる。

【0028】請求項5記載の発明では、複数の単位サーバは、それぞれ複数の単位サーバから構成された2以上のグループに分割されており、予め定められた複数の転送先はそれぞれ他のグループの単位サーバに設定されている。

【0029】すなわち請求項5記載の発明では、各単位サーバの有する第1のデータ記憶手段に記憶されたものと同一のデータは、その単位サーバの属する以外のグループの単位サーバの第2のデータ記憶手段に格納される。

【0030】

【発明の実施の形態】

【0031】

【実施例】図1は、本発明の一実施例におけるデータ格納システムの構成の概要を表わしたものである。このシステムでは、ネットワーク11を通じて第1～第4のディスク制御装置12～15が接続されている。第1のディスク制御装置12にはバス16を通じて通常の運用に用いられるディスク装置としての第1の通常用記憶装置17と、障害対策用に設けられたディスク装置としての第1の障害用記憶装置18が接続されている。同様に第2のディスク制御装置13には、バス21を通じて第2の通常用記憶装置22と第2の障害用記憶装置23が接続されている。また第3のディスク制御装置14にはバス24を通じて第3の通常用記憶装置25および第3の障害用記憶装置26が、第4のディスク制御装置15にはバス27を通じて第4の通常用記憶装置28および第4の障害用記憶装置29がそれぞれ接続されている。

【0032】ディスク装置への読み書きを制御するディスク制御装置とこれにバスを介して接続されている通常用記憶装置および障害用記憶装置をまとめて単位サーバと呼ぶ。また、複数の単位サーバがネットワークに接続されたものをクラスタと呼ぶ。ここでは、第1～第4の単位サーバ31～34によってクラスタが構成されている。図1に示したシステムでは通常用記憶装置および障

10

害用記憶装置としてハードディスク装置を用いている。このほか、フロッピーディスク装置、シリコンディスク装置などを用いることができる。また書き込み動作が必要なれば、CD-ROMを用いることも可能である。

【0033】各ディスク制御装置に接続されているバスは、SCSI (Small Computer System Interface) バスを用いている。これ以外にも、ファイバーチャネル等のシリアルバスあるいはATM(Asynchronous Transfer Mode)等のネットワークを用いることも可能である。クラスタを構成する単位サーバの数は3つ以上であれば良いが、本実施例では、4台の単位サーバがネットワークに接続されたクラスタを示してある。

【0034】第1の通常用記憶装置17の記憶内容と同一内容のデータは、これの属する第1の単位サーバ31以外の第2～第4の単位サーバ32～34に属する障害用記憶装置23、26、29に分散して格納されるようになっている。第2の通常用記憶装置22の記憶内容のコピーは、第2の単位サーバ32以外の単位サーバ31、33、34の障害用記憶装置18、26、29に分散して格納される。同様に第3の通常用記憶装置25の記憶内容のコピーは、第1、第2および第4の障害用記憶装置18、23、29に、また第4の通常用記憶装置28の記憶内容のコピーは、第1～第3の障害用記憶装置18、23、26にそれぞれ分散して格納される。このように、各通常用記憶装置の内容は、その通常用記憶装置の属する単位サーバ以外の単位サーバに属する障害用記憶装置に分散して格納される。

【0035】図2は、単位サーバの有するディスク制御装置の構成の概要を表わしたものである。ディスク制御装置41は、読み書きの制御の中枢的機能を果たすCPU(中央処理装置)42を備えている。CPU42にはバス43を介して各種回路装置が接続されている。このうちROM(リード・オンリ・メモリ)44は各種プログラムや固定的データの格納された読み出し専用メモリである。RAM(ランダム・アクセス・メモリ)45は、プログラムを実行する上で一時的に必要となる各種データを記憶するためのメモリである。ネットワーク制御装置46は、ネットワークとの間で各種のデータやコマンドの入出力をを行うための回路装置である。SCSIコントローラ47は、通常用記憶装置48および障害用記憶装置49との間でデータの転送を行うための制御回路である。SCSIコントローラ47から出力されているSCSIバス上に通常用記憶装置48および障害用記憶装置49は接続されている。

【0036】図3は、図1に示したデータ格納システムの各記憶装置の記憶内容の一例を模式的に表わしたものである。図1と同一部分には同一の符号を付してありそれらの説明を適宜省略する。ここでは、通常用記憶装置に格納されるデータのコピーは、他の3つの単位サーバの障害用記憶装置に分散して格納されるので、各通常用

50

記憶装置の記憶領域をコピーを格納する障害用記憶装置の数に対応して3つに分割している。すなわち、単位サーバの数（ディスク制御装置の数）より1つ少ない個数の分割記憶領域に分割してある。そして分割されたそれぞれに分割領域にその識別名称を割り当てている。

【0037】第1の通常用記憶装置17の第1の分割領域51には“D1-1”と、第2の分割領域52には“D1-1”と、また第3の分割領域53には“D1-3”的識別名称を付与している。また、第2の通常用記憶装置22の第1～第3の分割領域54～56には“D2-1”、“D2-2”“D2-3”的識別名称を与えている。第3の通常用記憶装置25の第1～第3の分割領域57～59には“D3-1”、“D3-2”“D3-3”を、第4の通常用記憶装置28の第1～第3の分割領域61～63には“D3-1”、“D3-2”“D3-3”をそれぞれ識別名称として割り当てている。同様に第1～第4の障害用記憶装置18、23、26、29の記憶領域もそれぞれ3つに分割されている。

【0038】第1の通常用記憶装置17にデータを書き込む場合、書き込み先の分割領域が“D1-1”（51）ならば、第1の通常用記憶装置17の領域“D1-1”に書き込むと同時に、ネットワーク11を通して同一のデータが第2のディスク制御装置13にも渡される。第2のディスク制御装置13は、受け取ったデータを第2の障害用記憶装置23の第1の分割領域67に書き込む。同様にして、第1の通常用記憶装置17の分割領域“D1-2”（52）に書き込むデータは、第3のディスク制御装置14に転送され、第3の障害用記憶装置26の第1の分割領域71に書き込まれる。第1の通常用記憶装置17の分割領域“D1-3”（53）に書き込むときは、同一のデータが第4の障害用記憶装置29の第1の分割領域74にも転送されて書き込まれる。

【0039】このようにして、第1の通常用記憶装置17に格納されるデータは、第1の通常用記憶装置17に書き込まれると同時に第2～第4の障害用記憶装置23、26、29に分散して格納される。図3では、各記憶装置の記憶領域をSCSIで用いられる論理ブロックを単位に分割した場合を示してある。通常用記憶装置および障害用記憶装置の記憶容量はそれぞれ同一であり、各記憶装置の記憶領域はそれぞれ“N”個の論理ブロックに分割されている。

【0040】このとき、 $N \geq 3n-1$ となる最大のnを選び、各論理ブロックに“0”～“3n-1”的番号を割り付けてある。そして、各記憶装置の論理ブロックの番号が“0”～“n-1”的範囲を第1の分割領域に、論理ブロックの番号が“n”～“2n-1”的範囲を第2の分割領域にそれぞれ対応させている。さらに論理ブロックの番号が“2n”～“3n-1”的範囲を第3の分割領域に対応させている。

【0041】第1の通常用記憶装置17の第1の分割領

域（51）“D1-1”に格納されるものと同一のデータは第2の障害用記憶装置23の第1の分割領域67に格納される。したがって、第1の通常用記憶装置17の“0”～“n-1”的論理ブロックに格納されるデータは、第2の障害用記憶装置23の“0”～“n-1”的論理ブロックにその複製が作成される。同様にして第1の通常用記憶装置17の第2の分割領域52としての“n”～“2n-1”的論理ブロックに格納されるデータは、第3の障害用記憶装置26の“0”～“n-1”的論理ブロックにその複製が作成されている。このように論理ブロック番号によってコピー先の障害用記憶装置およびその記憶領域を対応付けることができる。このような対応関係は、各単位サーバのディスク制御装置の有するROMあるいはRAMに登録される。

【0042】図3のようにデータの格納されているデータ格納システムにおいて、通常用記憶装置もしくはディスク制御装置に障害が生じた場合の動作を説明する。

【0043】図3に示したシステムでは、各通常用記憶装置のそれぞれの分割領域とそのコピー先となる障害用記憶装置およびその障害用記憶装置内における格納領域との対応関係が、他の単位サーバのディスク制御装置に予め通知されている。たとえば、第1の通常用記憶装置17の第1の分割領域51、すなわち“0”～“n-1”的論理ブロックのコピー先が第2の障害用記憶装置23の第1の分割領域67（“0”～“n-1”的論理ブロック）であることが他の単位サーバのディスク制御装置に通知されている。データの要求元となるホストコンピュータ77はネットワーク11につながれている。

【0044】ホストコンピュータ77には、ディスク制御装置12に障害が起きた場合はその代替としてディスク制御装置13と通信すること、またディスク制御装置13に障害が起きたときは代替としてディスク制御装置14と通信することが予め設定されている。さらに、ディスク制御装置14に障害が起きたときは代替としてディスク制御装置15と、ディスク制御装置15に障害が発生したときはその代替としてディスク制御装置12とそれぞれ通信することが予め設定されている。

【0045】各単位サーバのディスク制御装置は、他の単位サーバから送られてくる情報を基にして、自身のバスに接続されている障害用記憶装置の各記憶領域に格納されるデータのコピー元の通常用記憶装置を自身の有するRAMに記憶するようになっている。また、データの読み出しが要求されるホストコンピュータ77は、各単位サーバのディスク制御装置に障害が生じているか否かを示した情報をホストコンピュータ77の有するRAMに記憶する。また障害が起きたときにその代替となるディスク制御装置を内部のメモリに記憶するようになっている。

【0046】今、第1のディスク制御装置12に障害が起きたものとする。データの要求元となるホストコンピ

ユータ77は第1のディスク制御装置12の障害発生を検出した以後は、第1の通常用記憶装置17のデータを読み出す場合は、読み出要求を第1のディスク制御装置12の代替として設定されている第2のディスク制御装置13へ送る。読み出すべき領域は第1の通常用記憶装置17における論理ブロックの番号で表わされている。

【0047】第2のディスク制御装置13では、ホストコンピュータ77から送られてきた読み出要求から論理ブロック番号を計算する。計算された論理ブロック番号が“0”～“n-1”的範囲内であれば第2のディスク制御装置13のバスに接続されている第2の障害用記憶装置23に読み出すべきデータのコピーが格納されていると判別する。そこで、第2のディスク制御装置13は自身のバス21に接続されている第2の障害用記憶装置23から対応するデータを読み出し、ホストコンピュータ77にネットワーク11を通じて送信する。

【0048】一方、読み出しの要求された論理ブロックの番号が“n”～“2n-1”的範囲であれば第2のディスク制御装置13は第3のディスク制御装置14にネットワーク11を通じて読み出要求を転送する。第3のディスク制御装置14は、受信した読み出要求を基にして自身のバス24に接続されている第3の障害用記憶装置26から該当するデータを読み出しホストコンピュータ77にネットワーク11を介して送信する。同様に論理ブロック番号が“2n”～“3n-1”的範囲であれば、第2のディスク制御装置13は第4のディスク制御装置15に読み出要求を転送する。これにより読み出要求の転送された第4のディスク制御装置15によって第4の障害用記憶装置29から該当するデータが読み出されホストコンピュータ77に送られる。

【0049】通常用記憶装置に障害が発生したことの検出は次のようにして行われる。第1～第4のディスク制御装置12～15は自身のバスに接続されている通常用記憶装置に対してデータの読み出要求を出す。そして読み出要求の送出先の通常用記憶装置から一定時間内に応答の到来しないタイムアウトエラーが生じたり、応答が無いときはその通常用記憶装置に障害が発生したものと判別する。

【0050】単位サーバのディスク制御装置に障害が発生したことは次のようにして検出している。クラスタを構成している単位サーバのディスク制御装置は、次の単位サーバのディスク制御装置に対して一定時間毎に自己が正常に動作していることを表わした動作確認メッセージを送る。1つ手前のディスク制御装置から一定時間以上に渡って動作確認メッセージが届かないときは、2つ手前のディスク制御装置へ正常動作を確認するための問い合わせを行う。2つ手前のディスク制御装置から問い合わせに対する応答が無いときはさらにその1つ手前のディスク制御装置に動作確認の問い合わせを行う。このように応答が得られるまで、次々と遡って正常動作を確認

するための問い合わせを行う。そして問い合わせを始めたディスク制御装置から応答の帰ってきたディスク制御装置までの間のディスク制御装置に障害が生じたものと判別する。

【0051】障害用記憶装置へのデータの分散の仕方は、データを格納した通常用記憶装置の属している単位サーバ以外の単位サーバに属している障害用記憶装置に格納すれば良く、格納される順序やその領域については図3に示した例に限定されるものではない。

10 【0052】その一例として通常用記憶装置の記憶領域を物理的に分割し、それぞれの領域を他の単位サーバに属する障害用記憶装置に振り分けて割り当てることができる。また、通常用記憶装置に書き込むファイルを単位にして、障害用記憶装置を割り当ても良い。

【0053】論理ブロック番号を用いて障害用記憶装置を割り当てる場合、先に説明したように論理ブロックを所定数連続するように割り当てるほか、隣り合う論理ブロックを互いに異なる障害用記憶装置に割り当てることもできる。すなわち、SCSIで用いる論理ブロックを単位に障害用記憶装置を割り振る場合に、通常用記憶装置にデータを書き込んだ論理ブロックの番号を、分散する障害用記憶装置の数で割った余りの値を基準に割り振る。

20 【0054】たとえば、1つの通常用記憶装置の記憶内容を3つの障害用記憶装置に分散して格納する場合、論理ブロックの番号を“3”で割った余りを基準に格納すべき障害用記憶装置を割り振る。この場合、余りは“0”～“2”的3つの値をとるので、これらの値にそれぞれ障害用記憶装置を予め割り当てるにより、データをコピー先に論理ブロック単位で分散して格納することができる。

30 【0055】また、単位サーバに接続されている通常用記憶装置が複数台ある場合、それぞれの通常用記憶装置をこれまで説明した記憶装置の分割領域として扱うことも可能である。たとえば、第1～第4の単位サーバによってクラスタが構成されており、それぞれの単位サーバに第1～第3の通常用記憶装置が接続されているものとする。このとき第1の単位サーバの第1の通常用記憶装置の内容を第2の単位サーバの障害用記憶装置にコピーする。また、第1の単位サーバの第2の通常用記憶装置の内容を第3の単位サーバの障害用記憶装置にコピーし、第1の単位サーバの第3の通常用記憶装置の内容を第4の単位サーバの障害用記憶装置にコピーする。

40 【0056】第1の変形例

【0057】これまで説明した実施例では、各単位サーバの通常用記憶装置の記憶内容を他の全ての単位サーバの障害用記憶装置に分散して格納しているが、第1の変形例では、一部の単位サーバに分散して格納するようになっている。

50 【0058】図4は、本発明の第1の変形例におけるデ

ータ格納システムの構成の概要を表わしたものである。このシステムでは、第1～第6のディスク制御装置81～86がネットワーク87に接続されている。それぞれのディスク制御装置81～86には、バス91～96を介して通常用記憶装置101～106と障害用記憶装置111～116が接続されている。第1の変形例におけるデータ格納システムでは、これらを第1の部分クラスタ121と、第2の部分クラスタ122に分割している。それぞれの部分クラスタにおいて実施例と同様にデータが分散して格納される。すなわち、第1の部分クラスタ121では、第1のディスク制御装置81に接続されている通常用記憶装置101の内容のコピーは、同一の部分クラスタ121に属する第2および第3のディスク制御装置82、83に接続されている障害用記憶装置112、113に分散されて記憶される。

【0059】これにより、実施例ではクラスタ全体で1つの単位サーバの障害しか許されないのでに対し、第1の変形例では各部分クラスタ内で1つまでの単位サーバの障害に対応することが可能となる。もちろん、部分クラスタの数は2つに限らず、これ以上の数であっても同様の効果が得られる。

【0060】第2の変形例

【0061】図5は、第2の変形例におけるデータ格納システムの構成の概要を表わしたものである。図4と同一の部分には同一の符号を付してありそれらの説明を適宜省略する。第2の変形例も第1の変形例と同様に第1および第2の部分クラスタ121、122にクラスタが分割されている。第1の変形例では、各部分クラスタ内の通常用記憶装置の記憶内容のコピーは同じ部分クラスタ内の障害用記憶装置に分散されて格納されている。これに対して第2の変形例では、各部分クラスタに属する通常用記憶装置の記憶内容のコピーは、他の部分クラスタに属する障害用記憶装置に作成されるようになっている。

【0062】第1の部分クラスタ121に属する第1～第3の通常用記憶装置101～103で構成される第1の通常用記憶装置クラスタ131の記憶内容は、第2の部分クラスタ122に属する第4～第6の障害用記憶装置114～116で構成される第2の障害用記憶装置クラスタ132にコピーされる。同様に第2の各部分クラスタ122に属する第4～第6の通常用記憶装置104～106で構成される第2の通常用記憶装置クラスタ133の記憶内容は、第1の部分クラスタ121に属する第1～第3の障害用記憶装置111～113で構成される第1の障害用記憶装置クラスタ134にコピーされる。

【0063】このように1つの部分クラスタに属する通常用記憶装置の記憶内容を他の1つの部分クラスタに属する障害用記憶装置に分散して格納することにより、複数の部分クラスタから構成されるクラスタ内で複数の単

位サーバの故障に対応することが可能となる。図5に示したように部分クラスタの数が3つ以下の場合は、複数のクラスタで同時に障害が起きると回復できないが、部分クラスタの数が4つ以上の場合には、複数の単位サーバの障害に対応することができる。

【0064】図6は、4つの部分クラスタから構成されるデータ格納システムの概要を表わしたものである。このシステムでは、第1～第12のディスク制御装置141～153がネットワーク154に接続されている。それぞれのディスク制御装置141～153には、バス161～173を介して通常用記憶装置181～193と障害用記憶装置201～213が接続されている。ネットワーク154を介して構成されるクラスタは、第1～第4の部分クラスタ221～225に分割されている。

【0065】ここでは、第1の部分クラスタ221に属する通常用記憶装置181～183で構成された第1の通常用記憶装置クラスタ231の記憶内容は第2の部分クラスタ222に属する障害用記憶装置204～206で構成された第2の障害用記憶装置クラスタ232に分散されて格納される。また、第2の部分クラスタ222に属する通常用記憶装置184～186で構成された第2の通常用記憶装置クラスタ233の記憶内容は第3の部分クラスタ223に属する障害用記憶装置207～209で構成された第3の障害用記憶装置クラスタ234に分散されて格納される。

【0066】同様に第3の部分クラスタ223に属する第3の通常用記憶装置クラスタ235の記憶内容は、第4の部分クラスタ224に属する第4の障害用記憶装置クラスタ236に分散されて格納される。また、第4の部分クラスタ224に属する第4の通常用記憶装置クラスタ237の記憶内容は、第1の部分クラスタ221に属する第1の障害用記憶装置クラスタ238に分散されて格納される。このように各部分クラスタの通常用記憶装置クラスタの記憶内容のコピー先を次の部分クラスタの障害用記憶装置クラスタにすることで、制限はあるものの複数の部分クラスタにおいて障害が起きる場合に対応できる。すなわち、隣り合わない部分クラスタで同時に障害が起きた場合に対応可能となる。

【0067】たとえば、第1の部分クラスタ221と第3の部分クラスタ223で同時に障害が起きた場合に対応可能になる。この場合には、第1の部分クラスタ221の内容を第2の部分クラスタ222の障害用記憶装置クラスタ232から読み出し、第3の部分クラスタ223の内容を第4の部分クラスタ224の障害用記憶装置クラスタ236から読み出すことができる。同様に第2の部分クラスタ222と第4の部分クラスタ224で同時に障害が起きた場合には、第1および第3の部分クラスタ221、223の障害用記憶装置クラスタ238、234を用いて障害を回復することができる。

【0068】第3の変形例

【0069】図7は、第3の変形例におけるデータ格納システムの構成の概要を表わしたものである。第3の変形例では、部分クラスタのようなグループ分けを行わず、各単位サーバの通常用記憶装置の内容を他の任意数の単位サーバの障害用記憶装置に分散してコピーするようになっている。このシステムでは、第1～第9のディスク制御装置241～249がネットワーク251に接続されている。それぞれのディスク制御装置241～249には、バス261～269を介して通常用記憶装置271～279と障害用記憶装置281～289が接続されている。

【0070】各単位サーバの通常用記憶装置の記憶内容のコピーを格納する障害用記憶装置は、各単位サーバの通常用記憶装置毎に設定される。例えば、通常用記憶装置271に格納されているデータのコピー先として、障害用記憶装置282～284を割り当てている。各通常用記憶装置毎にばらばらに設定すると管理が複雑になるため、ここでは、通常用記憶装置の属する単位サーバの次の単位サーバから連続する3台の障害用記憶装置をコピー先としている。もちろんコピー先は連続する必要は無く、ばらばらであっても構わない。

【0071】以上説明した実施例および第1～第3の変形例では、通常用記憶装置と障害用記憶装置は物理的に別々の装置を用いたが、同一の記憶装置内を通常用記憶装置としての記憶領域と障害用記憶装置用の記憶領域に分割して用いることもできる。また、ディスク制御装置は、ワークステーションやホストコンピュータを用いることもできる。

【0072】

【発明の効果】このように請求項1記載の発明によれば、各単位サーバの第1のデータ記憶手段の記憶内容の複製が他の複数の単位サーバの第2のデータ記憶手段に分散して作成されるので、1つの単位サーバに障害が起きたときでも、他の単位サーバの負担が大幅に増大する事がない。また単位サーバの有する第1のデータ記憶手段に障害が生じた場合に限らず、第1の書き込み手段などの単位サーバの他の回路部分に障害が生じたときでも、障害の生じた単位サーバの第1のデータ記憶手段に記憶されているものと同一のデータを他の単位サーバから読み出すことができる。

【0073】また請求項2記載の発明によれば、ファイル単位に他の複数の単位サーバに分散しているので、データの分散や、分散されたデータの読み出しを容易に管理することができる。

【0074】さらに請求項3記載の発明によれば、第1のデータ記憶装置の記憶領域を複数に分割したブロックごとに同一のデータの転送先が割り振られているので、データの分散や、分散されたデータの読み出しを容易に管理することができる。

【0075】また請求項4記載の発明によれば、各単位

サーバの有する第1のデータ記憶手段に記憶されたものと同一のデータは、その単位サーバの属するグループ内における他の複数の単位サーバの第2のデータ記憶手段に分散された格納される。これにより、各グループ内で1つの単位サーバの障害をリカバることができ、データ格納システム全体として2以上の単位サーバの障害に対応することができる。

【0076】さらに請求項5記載の発明によれば、各単位サーバの有する第1のデータ記憶手段に記憶されたものと同一のデータは、その単位サーバの属する以外のグループの単位サーバの第2のデータ記憶手段に格納される。たとえば、4以上のグループに分ければ、2以上のグループの単位サーバの障害に対応することができる。

【図面の簡単な説明】

【図1】本発明の一実施例におけるデータ格納システムの構成の概要を表わしたシステム構成図である。

【図2】図1に示したディスク制御装置の回路構成の概要を表わしたブロック図である。

【図3】図1に示したデータ格納システムの各記憶装置の記憶内容の一例を模式的に表わした説明図である。

【図4】本発明の第1の変形例におけるデータ格納システムの構成の概要を表わしたシステム構成図である。

【図5】本発明の第2の変形例におけるデータ格納システムの構成の概要を表わしたシステム構成図である。

【図6】4つの部分クラスタから構成されるデータ格納システムの概要を表わしたシステム構成図である。

【図7】本発明の第3の変形例におけるデータ格納システムの構成の概要を表わしたシステム構成図である。

【図8】従来から使用されているミラー方式を用いたデータ格納システムの構成の概要を表わしたシステム構成図である。

【図9】従来から使用されているディスク制御装置の障害に対応することのできるデータ格納システムの構成の概要を表わしたシステム構成図である。

【図10】従来から使用されているディスク装置の障害およびディスク制御装置の障害の双方に対応することのできるバス構成の簡易なデータ格納システムの概要を表わしたシステム構成図である。

【符号の説明】

40	11、87、154、251 ネットワーク 12～15、41、81～86、141～153、24 1～249 ディスク制御装置 16、21、24、27、91～96、261～269 バス 17、22、25、28、48、101～106、18 1～193、271～279 通常用記憶装置 18、23、26、29、49、111～116、20 1～213、281～289 障害用記憶装置 31～34 単位サーバ 42 CPU
50	

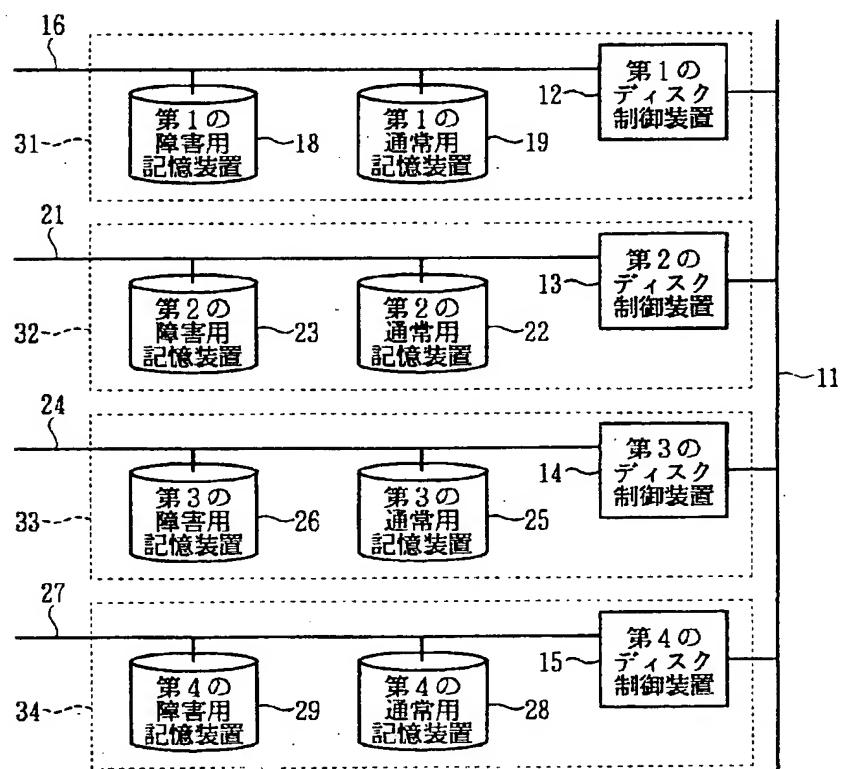
17

18

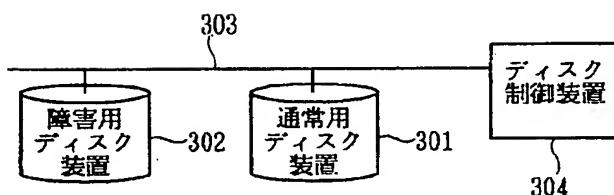
4 3 CPUバス
 4 4 ROM
 4 5 RAM
 4 6 ネットワーク制御装置
 4 7 SCSIコントローラ
 5 1～5 9、6 1～6 9、7 1～7 6 分割領域

7 7 ホストコンピュータ
 1 2 1、1 2 2、2 2 1～2 2 4 部分クラスタ
 1 3 1、1 3 3、2 3 1、2 3 3、2 3 5、2 3 7 通常用記憶装置クラスタ
 1 3 2、1 3 4、2 3 2、2 3 4、2 3 6、2 3 8 障害用記憶装置クラスタ

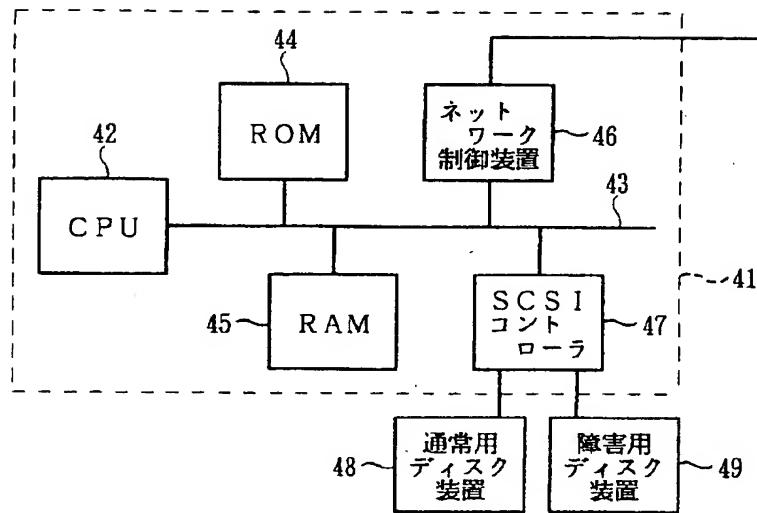
【図1】



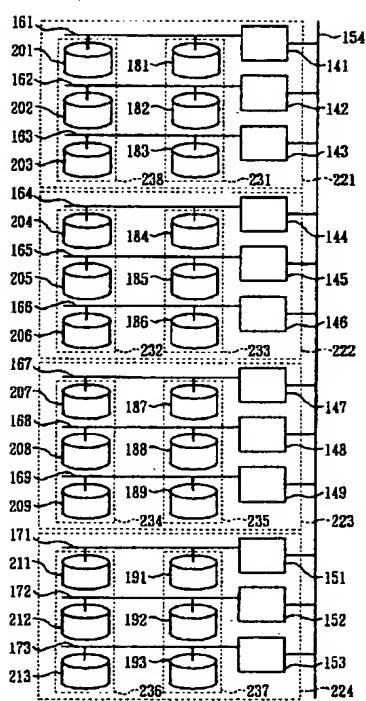
【図8】



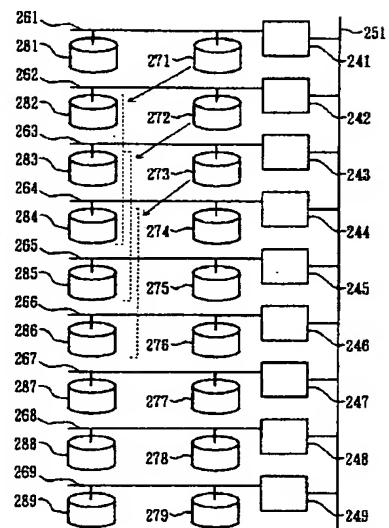
【図2】



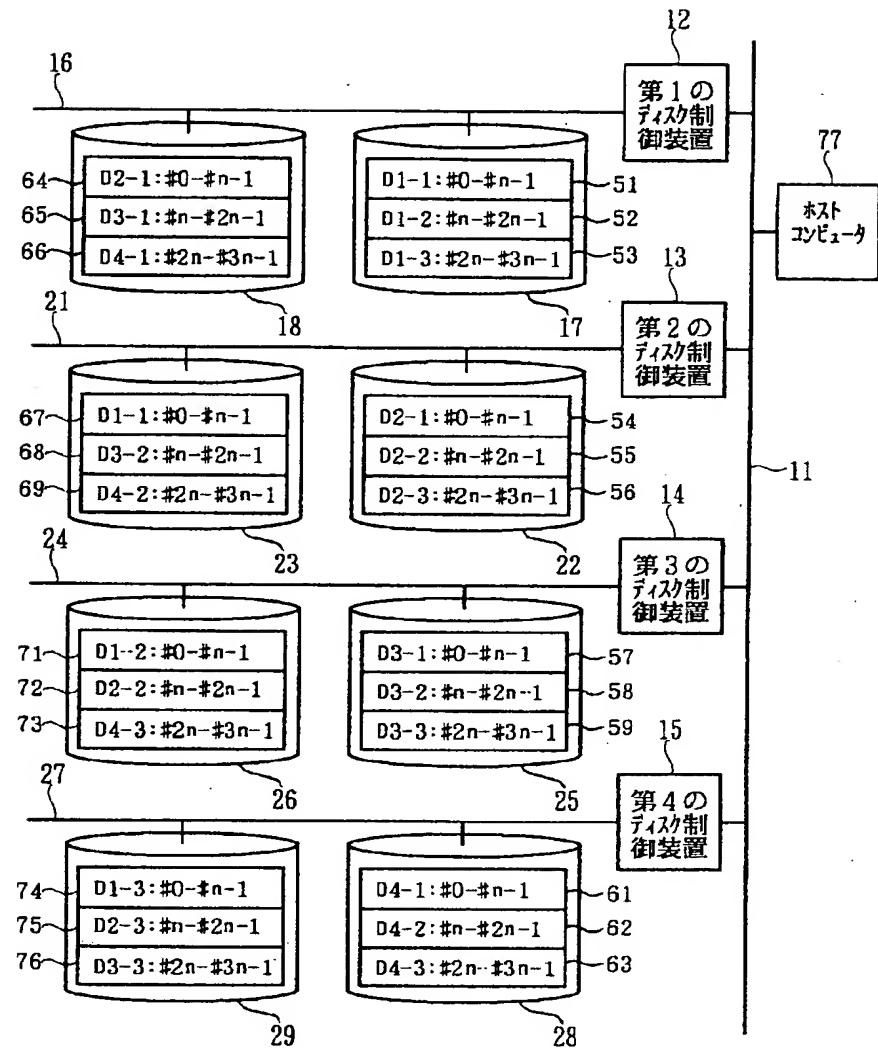
【図6】



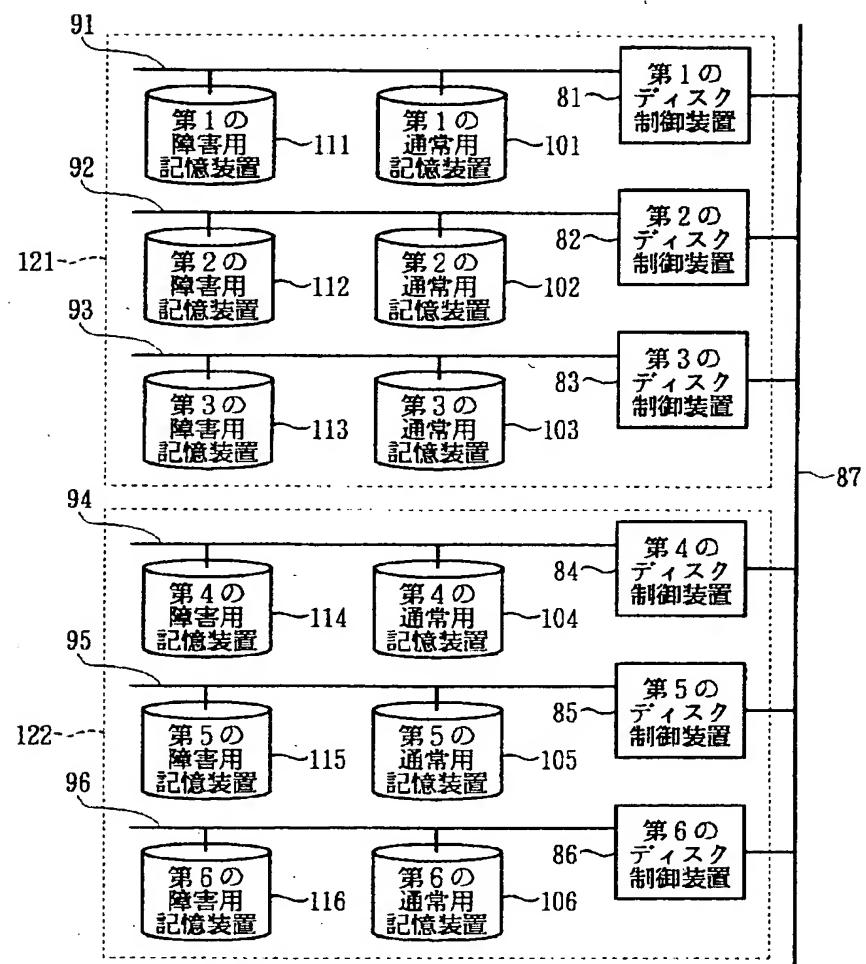
【図7】



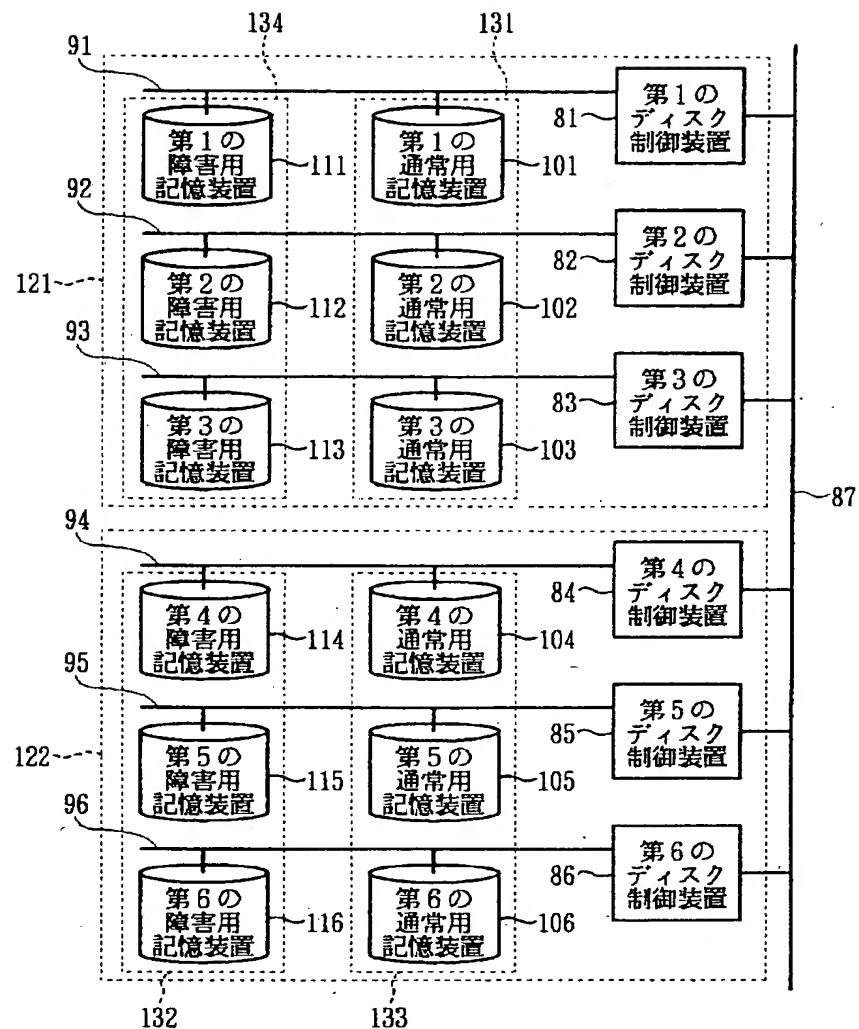
【図3】



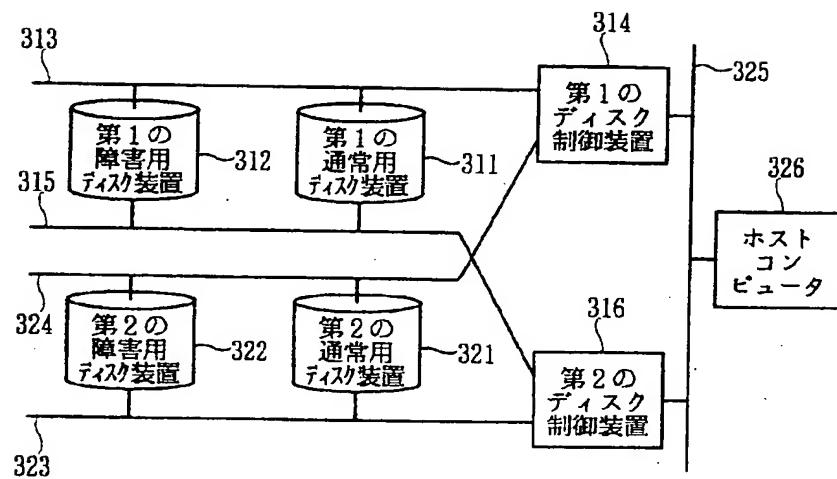
【図4】



[図5]



【図9】



【図10】

